DOCUMENT RESUME

ED 463 331 TM 033 763

AUTHOR Wang, Jianjun
TITLE Stochastic Models of Quality Control on Test Misgrading.
SPONS AGENCY California State Univ., Bakersfield.
PUB DATE 2002-04-00
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Testing; *Grading; Mathematical Models; *Measurement Techniques; Psychological Testing; *Quality Control; Scores; Test Results
IDENTIFIERS Binomials; Poisson Process; *Stochastic Analysis

ABSTRACT

Stochastic models are developed in this article to examine the rate of test misgrading in educational and psychological measurement. The estimation of inadvertent grading errors can serve as a basis for quality control in measurement. Limitations of traditional Poisson models have been reviewed to highlight the need to introduce new models using well established geometric and negative binomial distributions. Equations are developed for the use of geometric and negative binomial distributions in the study of test misgrading. In this study, the geometric process is developed from a single-grader scenario under a policy of zero tolerance for test misgrading. The negative binomial process seems appropriate for state or national assessment that involves more than one test grader. Results of this investigation can be used to ensure the number of misgraded events below a threshold k. Features of the quality control measures are discussed in this article in a context of local and national assessments. (Contains 22 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Running head: Test Misgrading

Stochastic Models of Quality Control on Test Misgrading*

ED 463 331

Jianjun Wang

Department of Teacher Education

California State University

9001 Stockdale Highway

Bakersfield, CA 93311-1099

E-mail: jwang@academic.csubak.edu

Phone: (661) 664-3048

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

J. Wang

1

TM033763

*Paper presented at the 2002 annual meeting of the American Educational Research Association (AERA), New Orleans, April 1-5. Support for this project was supported in part by the CSUB Faculty Teaching and Learning Center and a TNT grant in collaboration with Mr. Doug Gardner. The opinions expressed in this manuscript are solely those of the author, and do not necessarily represent these partners and organizations.

BEST COPY AVAILABLE

Stochastic Models of Quality Control on Test Misgrading

Abstract

Stochastic models are developed in this article to examine the rate of test misgrading in educational and psychological measurement. Limitations of traditional Poisson models have been reviewed to highlight the need of introducing new models using well established geometric and negative binomial distributions. Results of this investigation can be employed to ensure the number of misgraded events below a threshold k. Features of the quality control measures are discussed in this article in a context of local and national assessments.

Stochastic Models of Quality Control on Test Misgrading

In the last decade, essay items have been incorporated in major educational assessments, such as the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS) (Allen, Carlson, & Zelenak, 1999; Martin & Kelly, 1996). Meanwhile, classroom teachers are urged to use essay questions to complement multiple-choice items. Various responses generated from essay items demand a large amount of manpower in test grading. While no graders intend to make mistakes, accidental errors are likely to occur during the human operations (Wang, 1993). The purpose of this study is to examine the chance of test misgrading using appropriate models in statistics. The estimation of inadvertent grading errors can serve as a basis for quality control in educational and psychological measurements.

Literature Review

Statistical models have been sought to enhance quality control in various projects. In industrial statistics, quality control measures are adopted mainly to ensure the total number of inferior incidents below a threshold k. Bissssell (1970) reviewed,

> Incident counts form an important class of data, arising particularly in manufacturing processes and accident studies. ... It is often assumed that such events follow the Poisson Law. The assumptions of constant mean level and independence are often violated in practice. (p. 215)

In educational and psychological measurements, test misgrading can be treated as a specific type of incidents. In a classroom setting, Lyman (1998) noted that "Every teacher recognizes that grades are somewhat arbitrary and subjective" (p. 107). In a large-scale assessment, it is even more difficult to assume the same level of average performance among various graders. Accordingly, the assumption of a constant mean

performance level is often violated in small- and large-scale assessments, which makes the Poisson model unsuitable for most real-life applications (Rasch, 1980; Wang, 1993).

Whenever the assumption of Poisson distribution does not hold, statisticians tend to adopt alternative models to strengthen the quality control process. In particular, Johnson and Kotz (1969) pointed out, "The negative binomial distribution is very often a first choice as alternative when it is felt that a Poisson distribution might be inadequate" (p. 125). Edward and Gurland (1961) compared a class of distributions applicable to accidents, and reported that "the negative binomial gives an appreciably better fit than the Poisson distribution" (p. 504). Nonetheless, the negative binomial model has yet to be adopted in education to analyze test misgrading (Rasch, 1980; Wang, 1993).

In contrast, researchers in other fields have applied the negative binomial distribution on a wide range of topics. Barnwal and Paul (1988) reviewed applications of negative binomial models on count data, and noted that "Count data which follow the negative binomial distribution arise in numerous areas of biostatistics (Anscombe, 1949; Bliss & Fisher, 1953; Bliss & Owen, 1958; McCaughran & Arnold, 1976)" (p. 215). In military industries, much earlier applications have been made by Greenwood and Yule (1920). They reported that the negative binomial distribution gave a better fit than did the Poisson distribution to accidents in munitions factories in England during the First World War. Besides the count data, Ross and Preece (1985) added that "The negative binomial distribution is often appropriate for data for aggregated organism; it can arise from various different models (Anscombe, 1950, p. 360; Bliss, 1953, p. 185ff; Boswell & Patil, 1970; Freeman, 1980)" (p. 323). The various applications have resulted in different presentations of the negative binomial distribution. Consequently, as was noted by

Barnwal and Paul (1988), "Different authors have expressed the negative binomial distribution in different forms" (p. 215), causing substantial confusion in its applications.

Matloff (1988) further examined connections between the negative binomial distribution and other statistical models, and reported, "In spite of the fact the name of this family contains the word *binomial*, it is related more closely to the geometric family than to the binomial family" (p. 83). However, various alternative presentations have also been made in the statistical literature for the geometric distribution (e.g., Casella & Berger, 1990, p. 74 & p. 625). To avoid the distraction on the notation differences, geometric and negative binomial models have been introduced in this study to estimate the rate of test misgrading in education. Criteria of quality control have been considered to differentiate the models in various settings.

Stochastic Models

In a test scoring process, a contrast can be set to differentiate outcomes of correct-grading and misgrading. An event with dichotomous outcomes is typically modeled by a Bernoulli trial. For a well-designed test, the chance of misgrading (p) is not high. Quality control measures, such as arrangement of schedules for short breaking, can be introduced in the grading process to ensure that the number of misgraded cases is no larger than a specific level k. In practice, the grading process may continue until occurrence of the kth misgrading. By then, a break session can be scheduled to refresh the graders, and thus, help control the number of misgrading below level k.

To facilitate description of the stochastic model, one may define X to be the total number of successes before the kth misgrading, and f(x) to be the probability of obtaining exactly X successes. Accordingly, the total number of trials (X+k) depends on the

threshold level k and the number of correctly-graded cases (X) before reaching the threshold. Because the event of misgrading happens by accident, the number of correctly-graded cases (X) may vary among the graders. Given a level of the threshold k, the expected value of X can be employed to schedule break sessions before reaching a misgrading incident on the (X+k)th trial .

Geomatric Stochastic Process

Under a condition of zero tolerance, one may wish to schedule a break period for test graders before the first occurrence of misgrading. Using symbol s to represent successful test grading and m to represent the first misgrading, one may describe the stochastic process in the following chain of events:

$$\underbrace{ss \ldots\ldots s}_{X \text{ times}}\, m$$

This stochastic chain can occur in only one way, i.e., the grader has successfully graded test questions on the first X trials, and ends up with the first misgrading case on the (X+1)th trial. Therefore, the chance of obtaining exactly X successes before the first failure is:

$$f(x) = \underbrace{(1-p)(1-p) \ldots (1-p)}_{X \text{ times}}\, p$$

where p is the probability of misgrading in each trial.

This stochastic process can be described by a probability function for different values of X:

$$f(x) = (1-p)^x p; \qquad x = 0, 1, 2, \ldots . \qquad (1)$$

Equation (1) defines a geometric process "because the probabilities form a geometric series with a common ratio 1-p" (Kalbfleisch, 1979, p. 118). Since the chance of

misgrading in each trial (p) is less than 1, the total probability follows:

$$[1+(1-p)+(1-p)^2+\ldots]p=\frac{p}{1-(1-p)}=1$$

which confirms the feature of

$$\sum_{x=0}^{\infty}[f(x)]=1 \qquad (2)$$

While the expected value of X is an important index for setting break sessions in a test grading process, it is not so simple to derive E(X) using a definition formula $E(X) = \Sigma [X\ f(x)]$ (Casella & Berger, 1990). Fortunately, equation (2) provides a short cut for the statistical inference. Take derivatives on both sides of (2), and one may get

$$\frac{d}{dp}\sum_{x=0}^{\infty}[f(x)]=0 \qquad (3)$$

The left hand side can be adjusted as

$$\frac{d}{dp}\sum_{x=0}^{\infty}[f(x)]=\sum_{x=0}^{\infty}\frac{d}{dp}[f(x)]$$

$$=\sum_{x=0}^{\infty}\frac{d}{dp}[(1\text{-}p)^x\ p]$$

$$=\sum_{x=0}^{\infty}\ [(1\text{-}p)^x - x\ (1\text{-}p)^{x-1}\ p]$$

$$=\frac{1}{p}\sum_{x=0}^{\infty}(1-p)^x\,p-\frac{1}{1-p}\sum_{x=0}^{\infty}x(1-p)^x\,p]$$

$$=\frac{\sum_{x=0}^{\infty}f(x)}{p}-\frac{E(X)}{1-p}$$

Following equations (2) and (3), one may get

$$\frac{1}{p}-\frac{E(X)}{1-p}=0$$

Thus, the expected number of correctly graded cases is:

$$E(X) = \frac{1-p}{p} \tag{4}$$

Because a break session is arranged before occurrence of the first misgrading (i.e., k=1), the waiting time for the first misgrading is:

$$E(X+k) = E(X+1) = E(X) + 1 = \frac{1-p}{p} + 1 = \frac{1}{p} \tag{5}$$

Although the geometric distribution has been discussed in many books, some books presented the expected value in equation (4) (e.g., Bhat, 1984, p. 106; Casella & Berger, 1990, p. 74; Feller, 1957, p. 210; Port, 1994, p. 247) and others gave the result in equation (5) (e.g., Matloff, 1988, p. 83; Ewart, Ford, & Lin, 1974, p. 322; Draper, & Lawrence, 1970, p. 120). No authors clarified the variable difference between equations (4) and (5). In this study, the differentiation of X and (X+k) shows agreement of both presentations in the same stochastic process.

Results in equation (5) can be interpreted in a context of test misgrading. For a given test, if the chance of misgrading (p) is small, then the waiting time for a break session can be longer. In an extreme case, multiple-choice tests are graded by a machine which has p equal to zero in each trial. According to equation (5), the waiting time can be infinite, and thus, there is no need for a rest unless the grading machine is broken down.

A Negative Binomial Model

In a more complicated situation, a test is graded by a total number of n graders. For any grader i, the threshold k_i depends on the quality control requirement, and may take a value larger than one. To ensure the k_ith misgrading event occurring on the (k_i+X_i)th trial, the previous (k_i-1) misgrading events must already happen in the

preceding (k_i+X_i-1) trials. The probability of obtaining (k_i -1) misgrading cases on the first (k_i+X_i-1) trials is given by

$$f(x) = \binom{k_i + X_i - 1}{k_i - 1} p^{k_i - 1}(1-p)^{X_i} \qquad X_i = 0, 1, 2, \ldots \qquad (6)$$

where the probability of misgrading on trial (k_i+X_i) is p.

Equation (6) follows a negative binomial distribution (Feller, 1957). The probability generation function for the negative binomial distribution is $p^{k_i}[1-(1-p)t]^{-k_i}$. Because the overall quality control is based on cumulative performance of the n graders, the stochastic process involves n independent random variables, $X_1, \ldots X_n$. Thus, the quality control threshold k hinges on the distribution of $\sum_{i=1}^{n} X_i$. Fortunately, the probability generation function for $\sum_{i=1}^{n} X_i$ follows

$$P_{\sum_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} p^{k_i}[1-(1-p)t]^{-k_i} = p^{k}[1-(1-p)t]^{-k} \qquad (7)$$

where $k=k_1 + \ldots + k_n$. Based on uniqueness of the probability generating function (Port, 1994), $\sum_{i=1}^{n} X_i$ must have a negative binomial distribution with parameters k and p.

The expected value for $\sum_{i=1}^{n} X_i$ can be derived from the negative binomial distribution, and the result is k(1-p)/p (see Casella & Berger, 1990). Therefore, the waiting time before occurrence of the kth misgrading is:

$$E[(\sum_{i=1}^{n} X_i) + k] = \frac{k(1-p)}{p} + k = \frac{k}{p} \qquad (8)$$

In a comparison between (5) and (8), one may note that the geometric process can be

treated as a special case (k=1) of the negative binomial distribution. Based on the results in (8), the waiting time for a break period can be longer if the overall tolerance level k is higher and the chance of misgrading (p) is small.

In summary, partly due to differences in the notation choice, the well-established geometric and negative binomial distributions have yet to be used in models of test misgrading. In other fields, Johnson and Kotz (1969) have noted that "the negative binomial distribution is frequently used as a substitute for the Poisson distribution when it is doubtful whether the strict requirements, particularly independence, for a Poisson distribution will be satisfied" (p. 135). Thus, the geometric and negative binomial models provide alternative choices that are more flexible than the Poisson model in educational and psychological measurements.

Given the connection between geometric and negative binomial distributions, applications of these stochastic models hinge on characteristics of a specific setting. The geometric process is developed from a single-grader scenario under a policy of zero tolerance for test misgrading. Thus, the result in equation (5) may be more applicable in a local setting in which a teacher has been assigned to grade tests for an entire class. The negative binomial process, on the other hand, seems appropriate for state or national assessment that involves more than one test grader. In both cases, the waiting time for test misgrading has been derived from the corresponding stochastic processes. The results can be employed to schedule break periods to ensure the error of misgrading below a threshold k.

References

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). The NAEP 1996 technical report. Washington, DC: National Center for Education Statistics.

Barnwal, K. & Paul, S. R. (1988). Analysis of one-way layout of count data with negative binomial variation. Biometrika, 75, 215-222.

Bhat, U. N. (1984). Elements of applied stochastic processes. New York, NY: John Wiley & Sons.

Bissell, A. F. (1970). Analysis of data based on incident counts. The Statistician, 19 (3), 215-247.

Casella, G and Berger, R. L. (1990). Statistical inference. Pacific Grove, CA: Brooks.

Draper, N. R., & Lawrence, W. E. (1970). Probability: An introductory course. Chicago, IL: Markham.

Dunbar, S. B., et al. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, 4 (4), 289-303.

Edwards, C. B. & Gurland, J. (1961). A class of distributions applicable to accidents. Journal of the American Statistical Association, 56 (295), 503-517.

Ewart, P. J., Ford, J. S., & Lin, C. Y. (1974). Probability for statistical decision making. Englewood Cliffs, NJ: Prentice Hall.

Feller, W. (1957). An introduction to probability theory and its applications (2d ed.). New York, NY: John Wiley & Sons.

Greenwood, M., & Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the

occurrence of multiple attacks of disease or of repeated accidents. Journal of Royal Statistical Society, 83, 255-279.

Hinz, P. & Gurland, J. (1967). Simplied techniques for estimating parameters of some generalized Poisson distributions. Biometrika, 54, 555-566.

Johnson, N. L., & Kotz, S. (1969). Discrete distributions. New York, NY: Houghton Mifflin.

Kalbfleisch, J. G. (1979). Probability and statistical inference I. New York: Springer-Verlag.

Lyman, H. B. (1998). Test scores and what they mean. Boston, MA: Allyn & Bacon.

Martin, M. & Kelly, D. (1996). Third Iternational Mthematics and Science Study: Technical report. Chestnut Hill, MA: Boston College.

Matloff, N. S. (1988). Probability modeling and computer simulation. Boston, MA: PWS-KENT.

Parzen, E. (1962). Stochastic processes. San Francisco, CA: Holden-Day, Inc.

Port, S. C. (1994). Theoretical probability for applications. New York, NY: John Wiley & Sons.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago, IL: University of Chicago Press

Ross, G. J. S., & Preece, D. A. (1985). The negative binomial distribution. The Statistician, 34, 323-336.

Wang, J. (1993). Simple and hierarchical model for test misgrading. Educational and Psychological Measurement, 53, 597-603.